AAMRL-TR-88-035

USING DEPTH RECOVERY IN HUMANS (U)

THOMAS K. KUYK, Ph.D.

ICON CONSULTANTS
3541 GREAT OAK LANE
BIRMINGHAM, AL 35223

JULY 1988

FINAL REPORT FOR PERIOD AUGUST 1987 - JUNE 1988

Approved for public release; distribution is unlimited.

HARRY G. ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY
HUMAN SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6573

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for Public Release; Distribution is Unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AAMRL-TR-88-035
6a. NAME OF PERFORMING ORGANIZATION ICON Consultants 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION AAMRL/HEA
6c. ADDRESS (City, State, and ZIP Code) 3541 Great Oak Lane Birmingham, AL 35223	7b. ADDRESS (City, State, and ZIP Code) Wright-Patterson AFB, OH 45433-6573
8a. NAME OF FUNDING/SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F33615-87-C-0541
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO.
62202F	7184	26	14

11. TITLE (Include Security Classification)
Visual Depth Recovery in Humans (U)

12. PERSONAL AUTHOR(S)
Kuyk, Thomas K., Ph.D.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM AUG 87 TO Jun 88	1988 July 7	57

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Image Processing, Depth Perception, Bionics, Gabor Functions, Disparity, Stereopsis, Human Visual System
23	03		
12	09		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This program investigated the relevant information content contained in a physiologically based model of the human visual system with regard to the efficient extraction of depth through stereopsis. A computational system was developed using the Gabor representational scheme to model the spatial weighting functions of simple and complex cells known to exist in primate visual cortex. The algorithm was implemented in a foveated representation produced by a complex-logrithmic, conformal mapping of each image yielding the advantage of both high resolution and a wide field-of-view. The algorithm was tested by extracting depth arrays from random-dot stereograms. The results demonstrated that the depth mapping was accurate and in excellent agreement with the percept produced when human observers fused the same stereo images. The significance of this approach is that it models not a visual system but rather an experimentally verified model of the primate visual system. This is critical to the effective application of human information processing techniques to specialized or intelligent image processing systems.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Michael W. Haas	(513) 255-8893	AAMRL/HEA

DD Form 1473, JUN 86 *Previous editions are obsolete*

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

PREFACE

The work reported on herein was performed under contract number F33615-87-C-0541 as a part of the Small Business Innovative Research (SBIR) program. This report documents work which was initiated 19 August 1987 and proceeded through 1 June 1988.

This work was performed by a team of scientists and engineers that included Gregg Irvin, Ph.D., Thomas Kuyk, Ph.D., and Kenneth Urban from ICON Consultants and James Gaska, Ph..D, and Lowell Jacobsen, Ph.D. from FOVEA Consultants.

This contract was conducted for the Air Force, specifically the Human Engineering Division of the Armstrong Aerospace Medical Research Laboratory, under the technical direction of Mr Michael W. Haas.

TABLE OF CONTENTS

INTRODUCTION

Information about the three dimensional world is conveyed to our visual system as a pair of two dimensional light intensity arrays imaged upon the retinae of our eyes. Our central nervous system, through the process of stereopsis, is able to recover from these images an accurate description of the three dimensional structure of the environment. As observers, we experience the process of stereopsis as the perception of depth. Since the invention of the stereoscope 150 years ago, the fundamental processes underlying stereo vision have been intensely investigated, initially with psychophysical experimentation, and, in more recent times, with electrophysiological studies of the mammalian visual cortex.

Until recently, the profound difficulty of solving stereopsis was not fully appreciated, in part because for sighted individuals the process is fast and effortless. It therefore came as a surprise in the late 1960's when it was found that computational approaches to stereopsis require immense amounts of computation, and that the algorithms then employed were grossly inadequate for all but the most simplistic imaging situations. In the past twenty years both an increased appreciation and understanding of the nature of this complex information processing problem has developed. Recent research into the psychophysical and physiological basis of stereopsis has produced many important new findings. These findings motivated the development, during this project, of a promising new approach to computational stereopsis. This report describes the new approach, its current implementation, and experimental results, (using random dot stereograms), that demonstrate its feasibility. We first discuss the fundamentals of the stereovision problem and review biological and computational vision research that is pertinent to its solution.

PROBLEM IDENTIFICATION

It has become clear that a comprehensive understanding of the problem of stereopsis involves not only the understanding of the psychophysical and physiological basis of stereopsis, but also an understanding of a complex problem in the area of visual information processing theory as well. To fully understand a complex visual information processing system, it is first necessary to understand the nature of the visual task that the system is required to solve.

With regard to stereopsis, the images sensed by the two eyes provide two highly correlated views of a single scene. However, the stereoscopic images of objects at different distances from the observer differ according to the unique projective relation of each eye to the scene. The positional difference in the projection of a single point on an object's surface onto the two retinae is commonly referred to as retinal disparity. Conversely, a pair of points, one in each retinal image, are said to be corresponding points when their inverse projections correspond to a single point on an object's surface. The major task involved in stereopsis is to determine which pairs of image points correspond to single object points for <u>all</u> object points in the scene. This is commonly referred to as the correspondence problem.

The magnitude of the retinal disparity of corresponding points is directly proportional to the difference in depth between a point on an object's surface and the fixation point of the two eyes. Therefore the variations in disparity across the population of all corresponding points can be used to produce a relative depth map of the surfaces in a scene. Importantly, with regard to the process of stereopsis, the sole information base available in a pair of stereoscopic images is this

disparity component of corresponding retinal points.

A general procedure for computing a depth map from stereoscopic images is as follows: 1. determine the pairs of coordinates in the two eyes which are projections of the same visible surface point in the scene, ie. find all "corresponding" points, 2. measure the disparity between all corresponding points, and 3. use the disparity map together with knowledge of the sensor geometry to recover the 3-D structure of the imaged scene. The first task in the above procedure, namely solving the correspondence problem, has historically proven to be the most challenging. Once the point correspondence problem is solved, rendering an unequivocal correspondence between points in both images, the remaining two tasks involve relatively straightforward applications of projective geometry. The novel approach developed in this report describes a computational solution to the correspondence problem whose implementation is based on our current knowledge of the relevant psychophysics, physiology and information processing theory of stereopsis.

The pioneering psychophysical research by Julesz (1960), in which he employed random-dot stereograms to study human stereopsis, produced critical information regarding where, in the chain of visual processing events, stereopsis occurs. Before Julsez's research, it was often assumed that the problem of correspondence was solved by first recognizing objects and their components, and then seeking correspondence matches between recognized details in each image. Such a scheme places the correspondence process quite late in the chain of visual information processing events, certainly after complex monocular object recognition processes have taken place. The Julesz experiment demonstrated that this could not be the case. With random-dot stereograms, the individual monocular images contain no information about visible surfaces, texture gradients, perspective, illumination gradients or object boundaries. The

only information available is that of horizontal disparity. Yet, when viewing random-dot stereograms, normal observers can readily perceive surfaces in depth, and this percept of depth only occurs after one achieves fusion. Therefore, human stereopsis does not require an awareness of the features to match in either monocular view, and, the process of stereopsis can occur independent of monocular object recognition.

The importance of this observation is that it allows us to formulate the process of human stereopsis as the computation of disparity information from a pair of stereo images, without any need for monocular cues (Poggio & Poggio, 1984).

CORRESPONDENCE PRIMITIVES

If human stereopsis does not rely upon complex object recognition processes to generate correspondence primitives, then what other types of primitives (locally computed image descriptors) might be considered? It is easy to enumerate possibilities since almost any spatially local linear or nonlinear image functional, ranging from the raw gray-scale intensity values to highly processed nonlinear descriptors, could be used with some degree of success. Indeed, numerous primitives for correspondence matching have been evaluated, a few of which are discussed briefly below.

The most rudimentary information available to the human visual system is a pair of two-dimensional intensity arrays. If a point on a physical surface is imaged onto both retinae, then one might suppose that the corresponding retinal image points would have similar intensity values. However, attempts to solve correspondence based on correlating the intensity values at each array point have had little success (Sperling, 1970). The problems with such an approach are as follows. First, two corresponding image points can have quite different intensity

values due to the different vantage points (perspectives) of the two eyes. Second, the presence of sensor noise inevitably produces local intensity errors that make correspondence matching difficult within regions of gradual illumination changes. Finally, image point intensity matching is psychophysically questionable since, as shown by Julesz (1971), stereo image pairs which differ greatly in their average intensity are readily fused by human observers and produce a depth percept.

We know from Julsez that high level representations such as recognized objects are not used as the correspondence primitives, and from Sperling that the non-specific primitives of intensity values are insufficient. Therefore, there must be some intermediate level of representation, between these two possible extremes, that provides the primitives used in the computation of stereopsis by the human visual system.

In the past decade a great deal of psychophysical, physiological and computational research has explored intermediate level stereo-matching primitives. The class of primitives which has received the most attention to date are those which indicate the presence of oriented luminance discontinuities or "edges" in the image data (Marr & Poggio, 1976; Dey, 1975; Nelson, 1975 & 1977; Marr, Palm & Poggio, 1978; Grimson, 1981; Mayhew & Frisby, 1981; Arnold, 1982; Gennery, 1980; Baker, 1980 & 1982; Baker & Binford, 1981). Indeed, the labeling of discontinuities, or edges, has been an early step in many image processing applications and a great deal of engineering research has been dedicated to the development of fast and reliable edge detection methods (for a review see Davis, 1975). In recent applications of this approach, one or more differential operators are generally applied to a multi-resolution pyramid of images; the coordinates of maxima, minima, or zero crossings are labeled at each level of resolution. Such

an approach underlies, for example, the Marr & Poggio theory (1979) for stereopsis which specifically uses labeled zero-crossing points in a multi-resolution, Laplacian-of-a-Gaussian filtered image pyramid as stereo-matching primitives. However, major criticisms of such edge primitives are often cited.

First, an early step toward obtaining such primitives involves the application of differential operators. In practice, this involves convolving the image with a discrete-domain filter mask which approximates the desired differential operation. Unfortunately, differential operators enhance the deleterious effects of sensor noise by amplifying the high frequency content of the image. Such ill effects are mitigated to some extent by smoothing the resulting function (the role of the Gaussian filter in Marr & Poggios' theory) but the introduction of an integral operator (smoothing) seems to defy the justification for using local differential operators in the first place.

A second common criticism of edge primitives is based upon the observation that they are "detected" (a highly nonlinear operation) and inevitably information is lost at this stage. This criticism has, for example, been leveled at Marr & Poggios' theory of representation with zero crossings since one can readily devise stereoscopic images which produce a depth percept when fused by human observers, but which contain no zero crossings whatsoever (Mallot & Bulthoff, 1987). One is forced to concede therefore, that if zero-crossing primitives are employed in human stereovision, then other unknown types of primitives must be available in addition to compensate for the loss of information at the (zero-crossing) detection stage.

A final objection to the use of edge primitives in stereo matching is their sparse distribution in the image plane, necessitating, therefore, an interpolation process to obtain a

dense stereo depth estimation.

One can compensate for some of the above difficulties by computing a large population of linear (non-detected) differential primitives to form a feature vector which is associated with each point in the image array. A correspondence is then sought by seeking points with highly correlated sets of these multiple measures. Such an approach was evaluated by Kass (1983), who used many partial derivatives of the image that were smoothed via oriented filters of different sizes. This approach can readily lead to primitives which are often arbitrary, non-physiologic and numerous.

The above-mentioned shortcomings can be overcome by representational schemes which use integral, rather than differential operators, to obtain stereo-matching primitives. Such primitives include, for example, the coefficients in Fourier-like, or Gabor, image representations. Their notable shortcoming relative to differential primitives, namely computational expense, is rapidly diminishing as we continue to benefit from dramatic advances in signal processing hardware technology. It is our belief that primitives for stereo matching that are based upon integral operators show much greater promise than their differential counterparts in terms of developing versatile artificial visual systems. In fact, as discussed below, recent electrophysiological evidence <u>strongly</u> suggests that integral operators underlie the primitives employed for stereo matching in the human visual system.

ELECTROPHYSIOLOGY

Convergence of inputs from the two eyes first occurs in the primary visual cortex, referred to as V1. Barlow, Blakemore & Pettigrew (1967) were the first to demonstrate that receptive fields of some V1 neurons in the cat were located at non-corresponding zones in the retinae of the right and left eyes, and that these neurons responded strongly when a line stimulus to each eye was positioned on these disparate retinal zones. (The receptive field of a visual neuron is the area on the retina where visual stimulation can change the response rate of the neuron). Barlow et al. suggested that these disparity sensitive cells play a crucial role in stereoscopic depth perception because, with fixation, these cells would be selectively stimulated by stimuli at different relative depths. These findings in cat have since been replicated in macaque monkey (Poggio & Fisher 1977), an animal whose depth discrimination abilities are similar to those of man (Bough, 1970). Other work has been aimed at characterizing the response properties of these neurons by (1) providing quantitative descriptions of their monocular filtering properties and (2) delineating the manner in which these monocular fields are combined. These two topics will be discussed separately below.

MONOCULAR FILTERING PROPERTIES

It is generally accepted that neurons in the primary visual cortex can be classified into at least two major functional types, namely, simple and complex. Many studies have shown that simple cells linearly summate information across their receptive field. Thus a description of the spatial filtering properties of simple cells can be obtained by specifying their spatial weighting functions. Complex cells exhibit strong non-linear response properties. The two functional types will be discussed

separately.

SIMPLE CELLS

Several investigators have noted that the spatial weighting function of simple cells closely corresponds to the basis set of a representational scheme proposed by Gabor (1946) as eminently suitable for communications. In the original Gabor scheme a one-dimensional function is expanded in terms of odd-symmetric and even-symmetric signals defined as a Gaussian multiplied by a sinusoid and cosinusoid, respectively. These Gabor functions have the interesting property that their joint uncertainty, defined as the product of the spread of the signals in frequency and the spread of the signal in space, is equal to the theoretical minimum.

In 1980 Marcélja introduced the Gabor scheme to the vision research community and showed that the one dimensional weighting function of simple cells are well fitted by Gabor functions. Pollen & Ronner (1981) demonstrated that simple cells which are tuned to the same spatial frequency and orientation and have the same axis of symmetry often have phase offsets of 90 degrees in their response to drifting gratings. That is, simple cells demonstrate the quadrature phase relationship required of Gabor elementary signals. More recent experiments have examined the two-dimensional spatial weighting surfaces of simple cells and have found that they are well described by two-dimensional Gabor functions (Palmer, Jones & Muillkin, 1985). These studies suggest that the Gabor scheme provides a useful model of the early stages of the visual process.

Following Daugman (1985) we define the 2D Gabor function, depicted in Figures 1a,b, as:

$$f(x,y)=\exp\{-\pi[(x-x_o)^2a^2+(y-y_o)^2b^2]\}\exp\{-2\pi i[u_o(x-x_o)+v_o(y-y_o)]\},$$

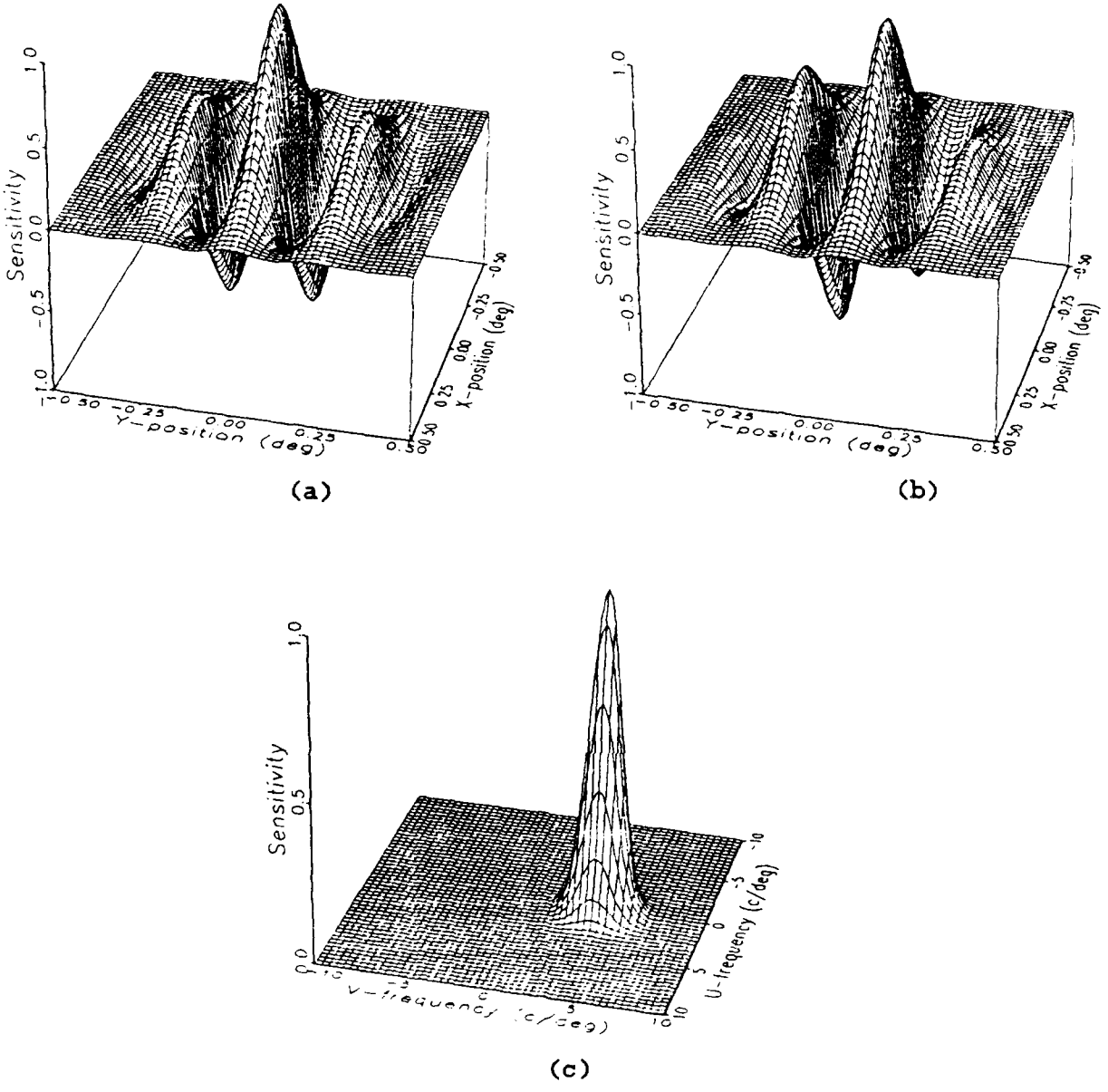


Fig. 1 The (a) real part and (b) imaginary part of a two-dimensional complex-valued Gabor filter, and (c) the filter's spatial-frequency magnitude spectrum.

with its 2D Fourier transform (Figure 1c) given by:

$$F(u,v)=\exp\{-\pi[(u-u_o)^2/a^2+(v-v_o)^2/b^2]\}\exp\{-2\pi i[x_o(u-u_o)+y_o(v-v_o)]\}. \quad (1)$$

There are eight parameters in the Gabor filter family given above:

Two spatial coordinates (x_o, y_o) which give the location of the filter in visual space.

Two modulation coordinates (u_o,v_o) which give the location of the filter in spatial frequency space.

The phase and amplitude which are specified by an assumed complex coefficient, $Ae^{i\theta}$, multiplying equation (1).

The width and length of the 2D elliptical Gaussian envelope (a and b), which are reciprocal in the two domains.

Although current evidence is not sufficient to provide a complete description of the parameters associated with the Gabor representation at the simple cell stage, partial answers are available. For example, the Fovea laboratory (Foster, Gaska, Nagler & Pollen, 1985) has measured the filtering properties of neurons in the primary visual cortex of the macaque monkey. The bandwidths of spatial frequency tuning curves were found to range from 1-3 octaves with a mean of about 1.6 octaves. The widths of the receptive fields, (related to the width of the elliptical Gaussian given in equation (1)), were inversely related to the optimal spatial frequency of the cells. Thus, the spatial frequency bandwidths (measured in octaves) were relatively constant at all spatial frequencies. Other

investigators (DeValois, Yund & Helper, 1982) have examined the orientation tuning bandwidths of V1 cells and found a median of about 40 degrees. In addition they found a positive correlation between orientation bandwidths and spatial frequency bandwidths. Daugman (1985) has analyzed this data and has shown that it is consistent with the behavior of a set of 2D Gabor filters with a field/width aspect ratio of 0.6. Thus, combining the information gleaned from physiological data with reasonable assumptions, such as a sampling density which is sufficient to avoid aliasing, we have developed a computational model, to be described shortly, which is similar to that found at the simple cell stage in the primary visual cortex of the monkey and, presumably, man.

COMPLEX CELLS

Several studies have shown that, when a stimulus whose contrast changes sinusoidally over time is presented within the receptive field of a complex cell, the cell's response is frequency doubled with respect to the input temporal frequency (Maffei & Fiorentini, 1973; Movshon, Thompson & Tolhurst, 1978). Furthermore, the complex cell's response to a drifting sine-wave grating is dominated by a large DC component with little or no modulated activity. Both of these non-linear behaviors can be accounted for by a model in which the complex cell squares the output from each member of a simple cell quadrature phase pair and then sums the result. These operations have the effect of computing the squared modulus of a complex Gabor coefficient. This model has recently been shown by Fovea (Pollen, Gaska, & Jacobson, 1987) to predict the responses of complex cells not only to single grating stimuli, discussed above, but also to the presentation of compound grating stimuli.

BINOCULAR CONVERGENCE

As earlier stated, the primary visual cortex (V1) is the first site in the visual pathway where inputs from both eyes converge on single cells; indeed, the majority of cells in V1 are responsive to stimulation in either eye. However, the rules of combination are different for the two cell types (simple and complex) discussed above.

This issue was addressed in an elegant series of experiments by Ohzawa & Freeman (1986 a,b) who dichoptically presented pairs of sinusoidal gratings with the same spatial frequency and drift velocity, but with relative phases (ie., disparities) which were varied over 360 degrees. The majority of binocularly driven simple cells (98%) responded with a modulated response whose frequency was equal to that of the input stimulus, and whose amplitude varied systematically with the phase offset between the dichoptically presented gratings. The responses of binocularly driven complex cells to the same stimulus regime fell into two equally prevalent response classes. Approximately 50% of the binocularly driven complex cells showed an unmodulated response whose amplitude varied with the phase offset between the dichoptically presented gratings; hence these cells are denoted as "phase-dependent complex cells." The remaining 50% of binocularly driven complex cells showed an unmodulated response that did not vary with the phase offset between the dichoptic gratings; these cells are referred to as "phase-independent complex cells."

The binocular cell responses observed by Ohzawa and Freeman can be summarized by three straightforward computational models. First, the responses of binocularly driven simple cells are described by a model in which the inputs from the two eyes are combined using a binocular, linear filter (Ohzawa & Freeman, 1986a). Second, the responses of the phase-dependent complex cells can be modeled by a non-linear (squared modulus) operation applied to the output of one or more binocular, linear filters

(possibly simple cells). Third, the responses of the phase-independent complex cells are predicted by a model which summates two or more nonlinear, monocular inputs which are themselves obtained by separately applying a nonlinear (squared modulus) operation to the outputs of monocular linear filters (Ohzawa & Freeman, 1986b). In all three cases, the spatial-frequency tuning characteristics of these cells are predicted if one assumes that the underlying linear filters have spatial profiles described by Gabor functions. The simulated responses of these three binocular cell types serve as the information base from which relative depth information is extracted in the system discussed below.

SYSTEM DESCRIPTION

A simplified outline of the stereo algorithm implemented in this project is shown in Figure 2. Given a pair of 2D stereo images, the algorithm is as follows:

0- Given a pair of stereoscopic images, then

for each disparity, d

1- Shift the left and right eye images horizontally by distances +d/2 and -d/2, respectively.

2- Both add and subtract these shifted images to produce a pair of interference images.

3- Compute two (complex-valued) space/spatial-frequency Gabor representations, one from the sum interference image, the other from the difference interference image.

4- Compute the square moduli (energy) of these two interference Gabor representations.

5- Add and subtract these energy representations to obtain two space/spatial-frequency representations which specify (a) the sum of the energy in the separate (monocular) right and left eye Gabor representations, and (b) the cross-correlation between the monocular right and left eye Gabor representations.

6- compute (a) a local 2D spatial energy representation by locally integrating the left/right eye energy sum over all

2D spatial frequencies, and (b) a local cross correlation function by integrating the left/right eye cross correlation over all 2D spatial frequencies.

7- compute a 2D spatial, normalized cross-correlation function by dividing the integrated left/right eye 2D cross correlation function by the integrated left/right eye 2D local combined energy function.

to obtain a 3-D normalized correlation function defined over one dimension of disparity and two dimensions of spatial position.

Then for each 2-D spatial position

8- let the local disparity estimate be that disparity for which the local 1-D normalized correlation function (of disparity) achieves its overall maximum.

to obtain a 2-D disparity function.

In a complete stereovision system, the 2-D disparity function obtained using the above algorithm would be further processed to obtain an absolute depth map. Though this additional processing is straightforward, it presumes knowledge of a particular sensor configuration that will vary depending on the application.

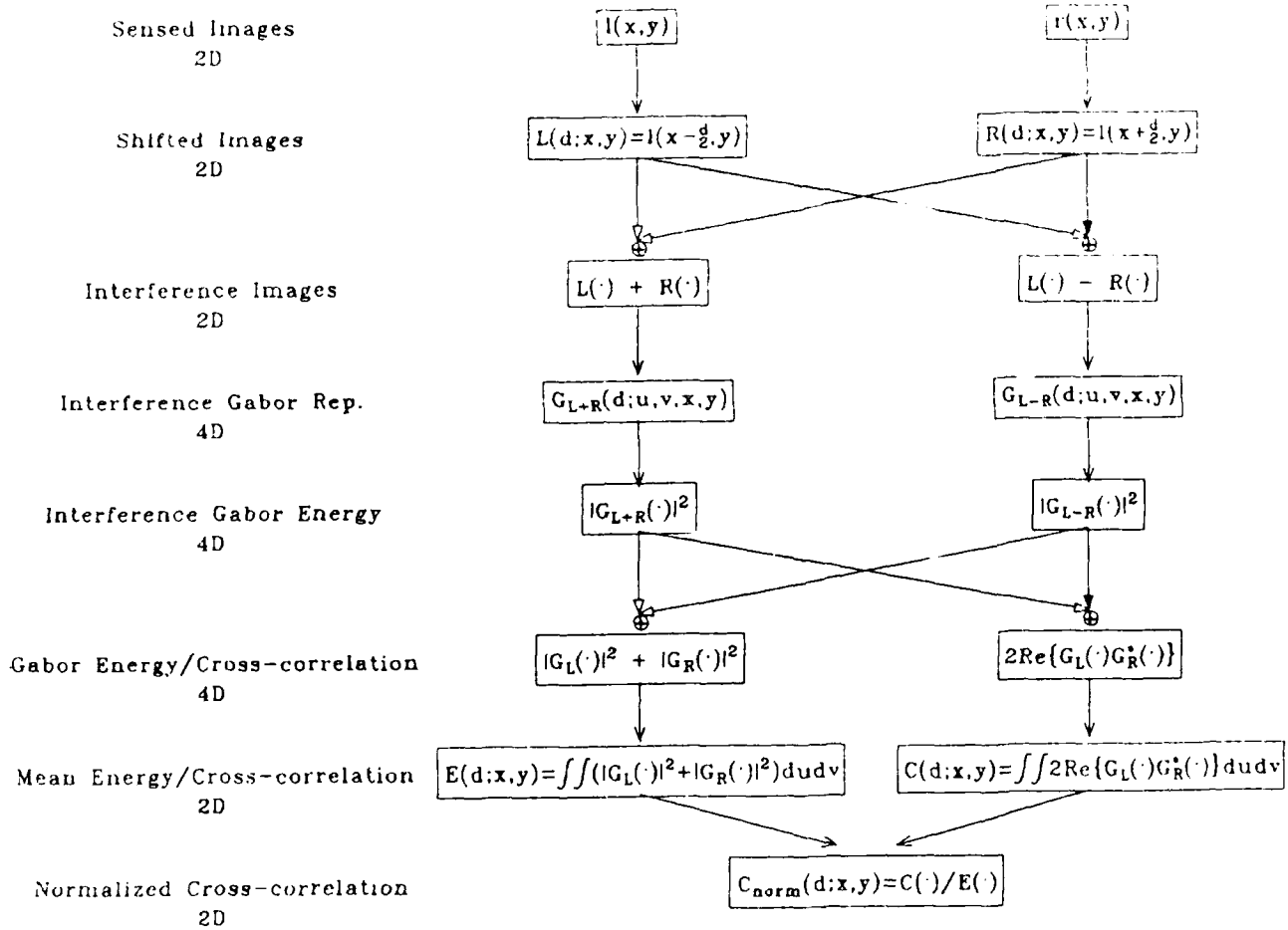


Fig. 2 Overview of the disparity estimation algorithm used in the implemented system.

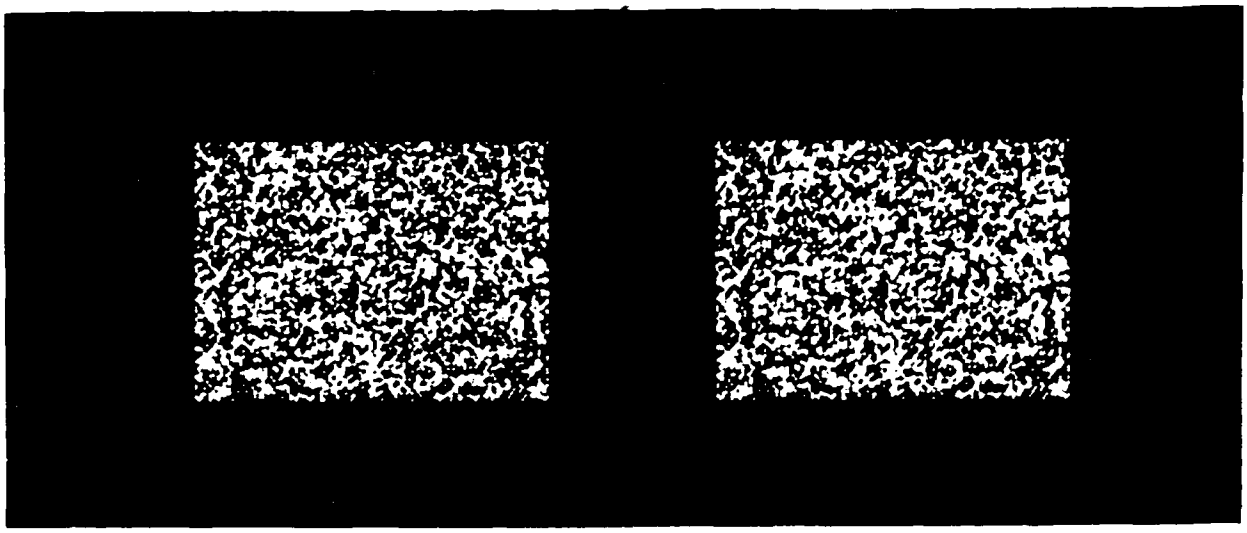
RESULTS

We now describe results obtained from a system which implements the disparity estimation algorithm presented in the previous section. For the particular example chosen, the input to the system consists of a random dot stereogram (Figures 3a,b) that was constructed as follows. The left eye image of the stereogram consists of random dots with an average density of 0.25 dots/pixel. The right eye image is identical to the left eye image except that a rectangular inner portion is shifted 3 pixels to the right, and the gap introduced by the shift is filled with an independent set of random dots. When viewing each image monocularly, a human observer sees a flat array of randomly placed dots; however, when fused, the observer perceives a central rectangular region in depth behind a rectangular foreground.

The system's computed solution to the stereogram in Figure 3 is depicted in Figure 4. The original left eye and right eye images have been colorized to indicate, for each eye's view, which image points were estimated to be projections from the near surface (red) versus the far surface (green). By fusing this colorized stereo pair, it is possible to observe where the computer's estimates differ from one's own depth percept. The computed solution is seen to agree very well with one's perceptual segregation of the surfaces in depth. In the remainder of this section, we describe in greater detail the actual computations performed to obtain the above result.

FOVEATION

The disparity estimation algorithm described in the previous section was formulated in terms of the cartesian coordinate frames of the original images, their displacements,



Left Eye Image Right Eye Image

Fig. 3 The left eye and right eye images of a random dot stereogram depicting a central rectangular region behind a rectangular foreground when fused divergently. Note that convergent viewing leads to a reversal in the depth relationship.

Left Eye Image Right Eye Image

Fig. 4 The system's computed solution to the stereogram in Fig. 3. The original left-eye and right-eye images have been colorized to indicate, for each eye's view, which image points were estimated to be projections from the near surface (red) versus the far surface (green). Note that by fusing this colorized stereo pair, one can observe where the computer estimates differ from one's own depth percept.

and their dual frequency domains. The actual implementation of the algorithm, which produced the results presented in this section, employs instead foveated representations of the left and right eye images. The definition and benefits of foveated representations are discussed briefly below.

The human visual system has long been known to employ an image sampling strategy which, despite limited computational resources, successfully provides both high spatial resolution and a wide field of view. These conflicting goals are achieved by providing a dense sampling of the retinal image within a small, centrally located receptive area, known as the fovea, and progressively diminished sampling with increasing distance from the fovea. The application of such foveated sampling schemes to machine vision systems would offer advantages over existing systems which provide uniform sampling resolution over a highly restricted visual field. In particular, a machine equipped with a foveated visual system would employ camera movements to selectively scrutinize the important targets in a scene with high resolution, while monitoring, (albeit with reduced resolution), the activity within its large field of view.

The foveated representation used in the current project is modeled after that of the human visual system. Each original image array is resampled onto a so-called polar exponential grid (Figure 5a) with the aid of a linear, space-variant, low-pass interpolating filter whose upper cutoff frequency drops with increasing eccentricity (Figure 5c). This results in a complex-logarithmic, conformal mapping (Figure 5b) of each image, where the mapping is performed with respect to a spatial origin within the central foveal region. The resulting conformally mapped images have orthogonal dimensions corresponding to the polar orientation and log-of-polar radius of points in the original image plane. Such a complex logarithmic conformal mapping has been shown to approximate the actual mapping of the retinal

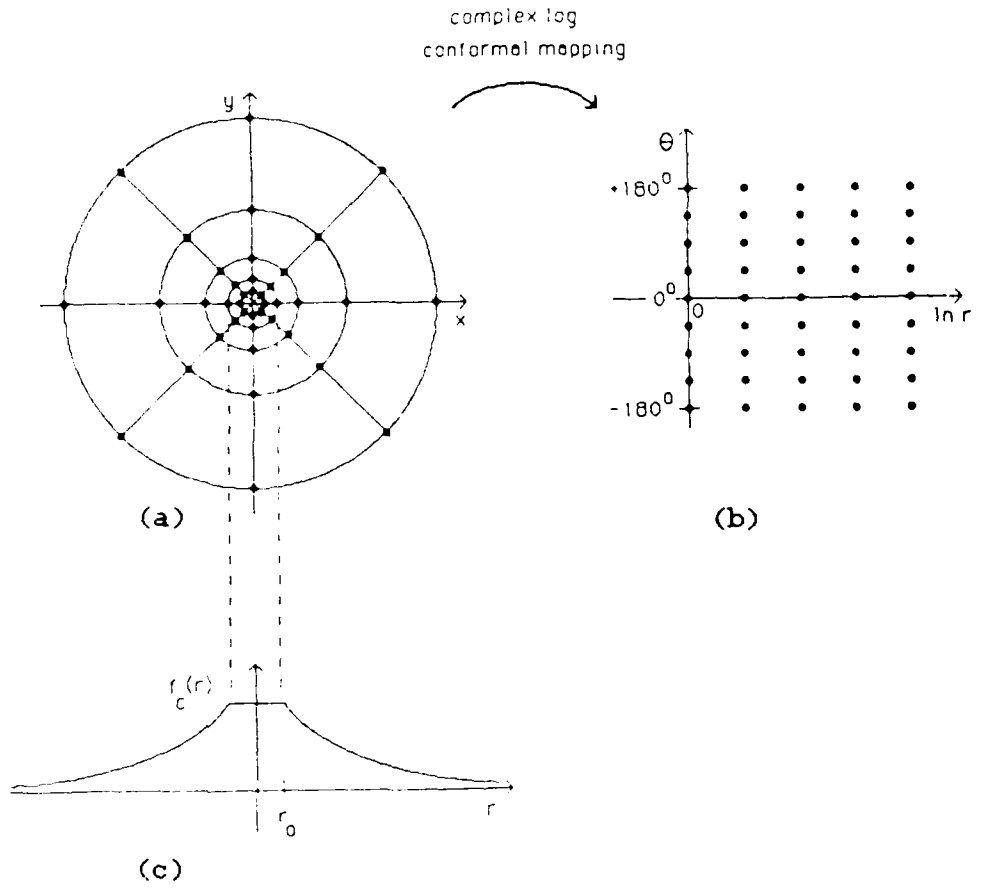


Fig. 5 The space-variant (foveated) image mapping used by the implemented system. (a) Sampling the sensor plane with a polar exponential grid (PEG), (b) remapping the sensor plane sample points using the complex logarithmic conformal mapping, (c) the high-frequency cutoff of the system as a function of visual field eccentricity.

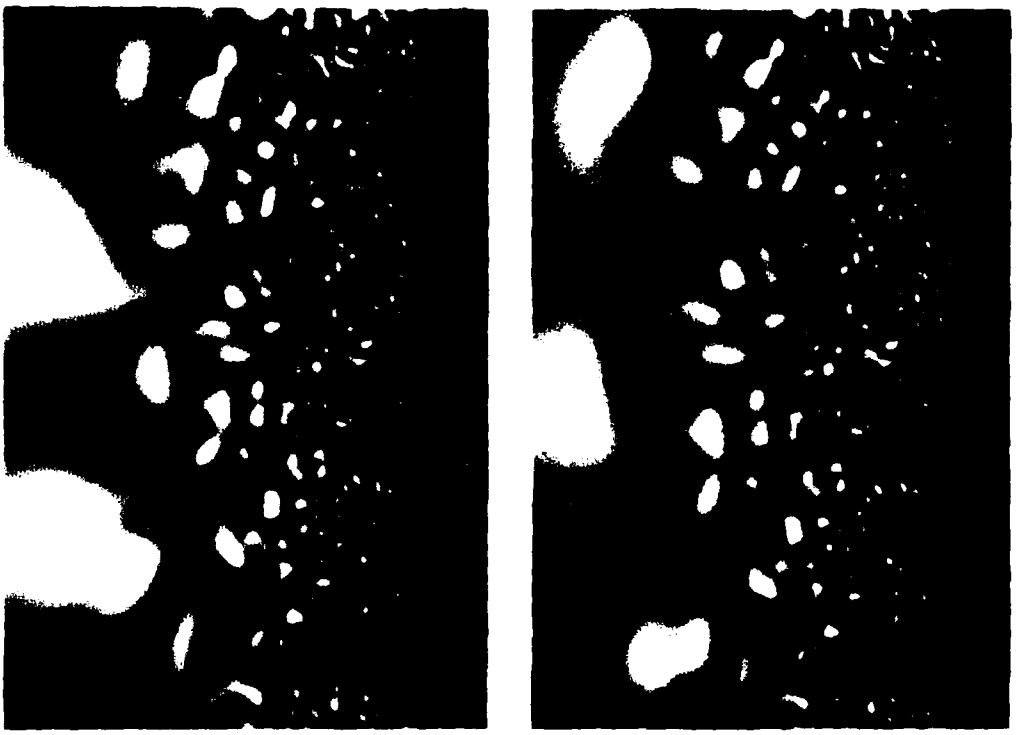
surface onto the surface of the visual cortex (Schwartz,1977).

The effect of this mapping on the separate left and right eye images of the random dot stereogram in Figure 3 is illustrated in Figure 6. The vertical axes on the mapped images specify the polar angle (0 degrees at the bottom), and the horizontal axes specify the log of the radial distance from the fovea. The lower edge of each conformally mapped image therefore corresponds to a line in the original cartesian image (Figure 3) which extends from near the image center horizontally to its right edge. Note that the mapping expands the representation of the central part of the visual field relative to that of the periphery. This is just as expected from a sampling scheme which is dense near the fovea and sparse in the periphery (see Figure 5a).

GABOR REPRESENTATION

In the next stage of processing, the system produces pairs of conformally mapped Gabor representations. This is done by convolving the conformally mapped images with specially modified complex-valued Gabor filters. These filters reflect the complex logarithmic variable transformation under the defining space-variant Gabor convolution integral. As a side benefit, the domain mapping converts the space-variant convolution integral into a space-invariant convolution integral, reducing therefore the complexity and computational expense of the Gabor filtering. The filtering at this stage produces a pair of discretely sampled 4D Gabor representations whose space and spatial-frequency domains are both conformally mapped.

Examples of 2D sections through such 4D complex-valued Gabor representations are presented in Figure 7. The images in the top row depict the real part of the representation and those in the bottom row depict the imaginary part of the



(a) (b)

Fig. 6 Conformally mapped images. (a) and (b) show respectively the conformally mapped versions of the left eye and right eye images in Fig. 3. In each case, conformal mapping was performed with respect to an origin at the point of visual fixation (the respective image centers).

representation. The images in the left column were computed from the conformally mapped left eye image (Figure 6a), and the right column images from the conformally mapped right eye image (Figure 6b). These sections through the monocular Gabor representations reflect the disparity relationship of the original images. At low eccentricities, (where the original images are disparate by 3 pixels within a rectangular region), the left and right eye Gabor functions are quite different as can be seen by inspection of Figure 7 (leftward regions). Conversely, at larger eccentricies, (where the original images have zero disparity), the left and right eye Gabor functions are indistinguishable (rightward regions). Such differences/similarities of left and right eye Gabor representations can be quantitatively measured using a cross-correlation process such as that defined in Figure 2.

Although Figure 7 depicts sections through monocular Gabor representations, in the actual disparity estimation system (Figure 2), one computes binocular Gabor representations obtained from the sum and difference of horizontally shifted left and right eye images. The computation of such a binocular Gabor representation is consistent with the electrophysiological data discussed earlier which indicates that the majority of simple cells linearly filter input from both eyes. Subsequent transformations of the Gabor representations of these sum and difference interference images lead to correlation results that are equivalent to those obtained by simply cross-correlating monocular Gabor representations of shifted left and right eye images.

CROSS-CORRELATION

A normalized cross-correlation between a pair of Gabor representations is computed for each of numerous disparity channels. The computations in each disparity channel differ

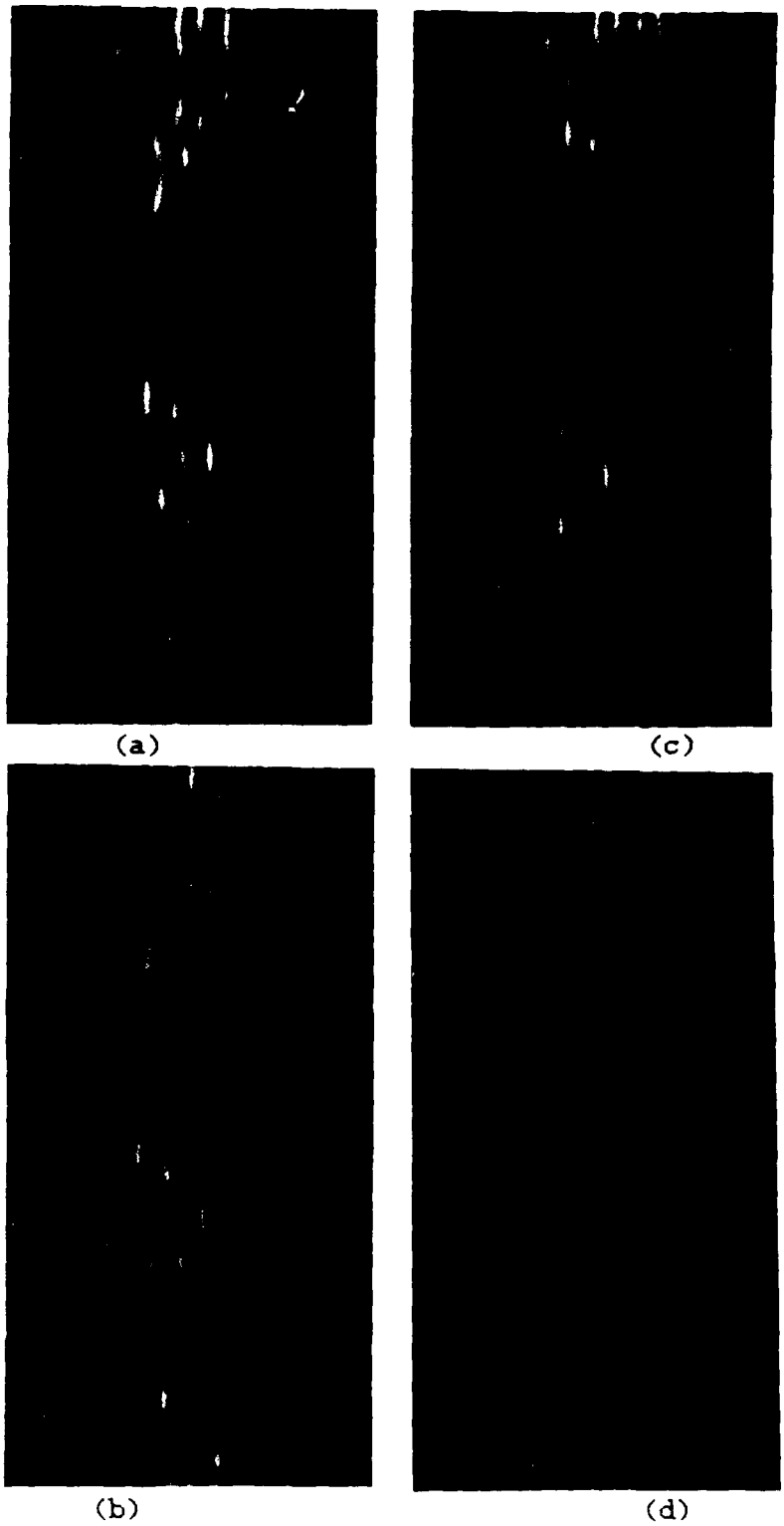


Fig. 7 Complex-valued Gabor spectra. (a) The real part, and (b) the imaginary part of a 2D section through the conformally mapped 4D Gabor spectrum obtained by filtering the conformally mapped left eye image in Fig. 6a; (b) and (d) show the related real and imaginary results obtained by filtering the conformally mapped right eye image in Fig. 6.

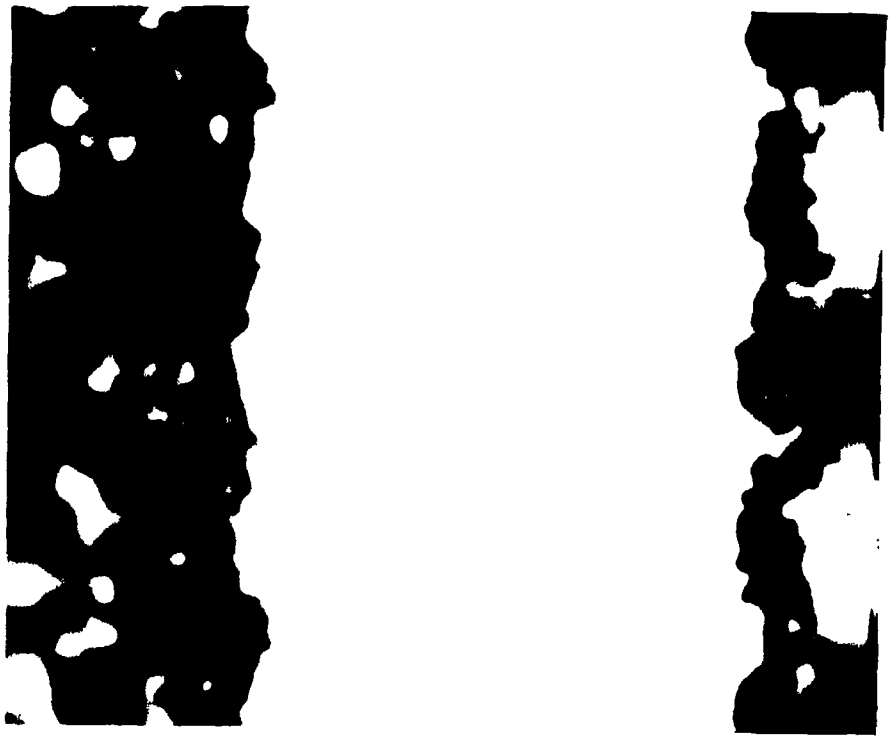
only with regard to the relative horizontal shifts applied to the right and left eye images as indicated by the "shifted images" stage in Figure 2. The output of all disparity channels can be combined to produce a discretely sampled, normalized correlation function defined over three dimensions; namely, two dimensions which specify position in a cyclopean visual field, and a third dimension of binocular disparity.

The outputs of two such disparity channels are shown in Figure 8, which depicts the normalized cross-correlation from the 0 and 3-pixel disparity channels. The correlation functions only take on values between -1 and +1. Bright and dark regions signify respectively positive and negative correlations. Note that the 0 disparity channel shows a strong positive correlation at larger eccentricities (rightward) and negative correlation throughout most of the small eccentricity region (leftward). The opposite is true for the 3 pixel disparity channel.

DEPTH MAP

Figure 9 shows the result obtained after processing the two correlation functions in Figure 8 to produce a 3D binary depth map. For each 2D position in the cyclopean visual field, the 3D map is bright only at the unique disparity which produced the maximum correlation value across all disparity channels. In the particular example shown, only two disparity channels (0 and 3 pixels) were used. Hence all points in the visual field were assigned by the algorithm to one of these two possible disparities.

Finally, to simplify the presentation of the solution, the conformally mapped disparity estimate of Figure 9 was remapped onto the original cartesian image domains of the two eyes to produce the result shown in Figure 4. Recall that Figure 4 indicates, for each eye's view, which image points were



(a) (b)

Fig. 8 Normalized cross correlation. (a) The normalized cross correlation obtained from the 0 disparity channel, and (b) the normalized cross correlation result for the 3 pixel uncrossed disparity channel. Bright regions denote high correlation.

(a) (b)

Fig. 9 Depth map. (a) A binary map of points in the cyclopean visual field that are estimated to have zero disparity (bright points), and (b) points estimated to have a disparity of 3 pixels (bright points).

estimated to be projections from the near surface (red) versus the far surface (green). By visually comparing the fused depth percept with the colorized zones, one observes that the solution is in good agreement with the perceptual segregation of the surfaces in depth.

FUTURE WORK

ENHANCED DISPARITY ESTIMATION FOR NATURAL IMAGES

In the example above, only two disparity channels were used to demonstrate the new algorithm with a simple pair of test images. However, when processing natural images, one would employ numerous disparity channels to obtain estimates of the continuously varying depth of surfaces in real-world scenes. An immediate objective in follow-on work would be to use a sufficient number of disparity channels to demonstrate the effectiveness of the algorithm on natural stereoscopic imagery.

The results shown in the previous section were obtained by pooling over Gabor filtering mechanisms of many orientations, but only a single frequency scale, at each spatial position. Because the random dot stereogram used in the example has a wide-band spectrum, we were able to obtain accurate results. However, when processing natural imagery, it will be necessary to pool over the full complement of Gabor filter orientations and scales. Indeed, the disparity estimation algorithm shown in Figure 2 is designed to pool over multiple orientation and frequency scales at each position. We have already implementated the filtering necessary to produce a Gabor representation at multiple scales, and future work should be directed at enhancing the current system to pool over this full Gabor representation.

Finally, rather than using a peak detection procedure to find the best estimate of disparity at each point in the cyclopean visual field, one might instead employ a more sophisticated procedure to estimate the disparity at each point. For example, the disparity channel correlations at each point in

the cyclopean visual field could be interpolated in order to estimate local disparity with finer resolution than that associated with the interchannel disparity sampling. With such an approach, one might be able to duplicate, (with a reasonable number of disparity channels), stereoscopic hyperacuity, i.e. the human ability to detect disparity differences as small as a few seconds of arc.

ABSOLUTE DEPTH ESTIMATION

From binocular disparities the human visual system recovers the three-dimensional shape of objects and computes estimates of their absolute distance relative to the observer as well as to other objects. Utilizing the information provided by horizontal disparity it is only possible to recover the relative depths of surfaces. Knowledge of the convergence angles as well as interocular separation is necessary to recover estimates of absolute depth from horizontal disparity information. In biological systems, the convergence angle of the two eyes is thought to be encoded by a signal of extraretinal origin. This signal could provide the information necessary to transform the (relative) disparity map into absolute depth information (Foley, 1980). However the precision of such a system and its effective availability are unknown (Cormack & Menenedez, 1983).

In an artificial system, the distance between the receiving elements (eyes) and their convergence angle can be precisely known. Though these parameters are fixed in simple stereoscopic systems, a truly versatile system would be able to vary these parameters in order to derive accurate depth information over a wide variety of target ranges. A major goal of future work is to unite the disparity estimation algorithm developed in this project with such a dynamic, convergent stereoscopic imaging system.

APPLICATIONS

The novel approach suggested in the technical portions of this proposal should be useful in a wide variety of tasks which require or would be enhanced by stereoscopic vision. Application of the technique would provide cogent depth information about the environment within which the system is operating.

The alternate paradigms proposed for machine use of stereoscopic depth information have been shown to have far less efficacy and accuracy, but the innovative approach suggested in this report should prove effective in its representation of depth with real world images.

The algorithm implementation outlined in the technical portions of this report has application far beyond the immediate scope of the project. The primacy of stereo vision (by virtue of its low-level position in the human visual information processing hierarchy) indicates it is an important requisite skill for operating within complex visual environments. Thus, it would be desirable to implement such visual processes in machine systems. The nature of the algorithms outlined in this report make it possible to implement stereo vision without artificial intelligence, i.e. it is a purely knowledge-free, computationally-based process. The technique will permit the application of human-analogue stereopsis to such diverse areas as remotely-piloted vehicles, autonomous manufacturing distribution vehicles, telepresence-based systems, image interpretation, and self-guiding vehicles typically envisioned by some military designers.

The requirements of each of these applications are quite varied, yet the flexibility and adaptive nature of the technique

make it possible to tailor the algorithm to a variety of disparate uses. Almost any type of situation requiring machine vision could benefit from this technique. The following section describes some of the unique requirements for possible operating environments.

POSSIBLE NEAR-TERM APPLICATIONS

THREE-DIMENSIONAL DISPLAYS

One area of application for human-based stereopsis is the development of three-dimensional displays. As equipment becomes more complex, the design and implementation of effective methods for humans to operate and interpret system states increase in difficulty. One method of combating this increased complexity is to provide the operator with displays which capitalize on his abilities. Displays which simulate objects or symbology in three-dimensional space enable the operator to more easily deal with large amounts of data in graphic rather than textual form. The development of such human-analogue stereopsis shows great promise both for the development and evaluation of future human-computer interfaces.

One area which could benefit from the implementation and diagnostic uses of the stereopsis algorithms outlined in this paper is the virtual cockpit. In many cases, present three-dimensional displays are less than optimal in their presentation of some information. Although the suboptimal presentation is recognized (often in the form of visual complaints from the user), the difficulty lies in quantifying the deficiencies. For example, there is only a limited range of disparity gradients over which a human subject can fuse two images. A display generator might produce gradients for a three-dimensional display just at the limit of the subject's ability to fuse the image, or could introduce subliminal drifts in disparity values.

Although the user might not notice the problem, the conflict within the visual system might be sufficient to produce visual discomfort. The algorithms described in this report make it easier to quantify and correct problems in the presentation through objective analysis of the imagery. Furthermore, application of the techniques in the development of the display imagery would insure display parameters within the human operating range. In effect, such development would tend to produce displays already optimized for operator performance. The Armstrong Aerospace Medical Research Laboratory and NASA are two organizations deeply involved in the development of such three-dimensional displays; they could benefit from applications which would permit detailed analysis of their alternate visual realities.

FUTURE APPLICATIONS

Future applications for the stereopsis algorithm proposed in this report rely on real-time analysis of the visual environment. The proposed technique is not only computationally efficient, but human based. This is significant because computations are only made within the human window of visibility. High spatial frequencies outside of the window of visibility (e.g. above the human's high spatial frequency cutoff) are not computed because it is unnecessary for adequate performance, and is too expensive computationally. The foveated nature of the system means high-resolution computation time is not wasted on the periphery, where humans don't use it. By virtue of its human-based design, this means engineers can design to the needs of the system, being precise only where it is important. By definition, any processing scheme developed with the system, when implemented on devices (displays) for humans, will have been built only to the requirements of necessary system. This makes system design very efficient and cost effective.

A brief summary of some systems requiring computation "on-the-fly" are outlined below.

REMOTELY-PILOTED VEHICLES AND AUTONOMOUS VEHICLES

Remotely-piloted vehicles (RPVs) have seen increasing use in both military and civil applications. In general, these are untethered vehicles which relay sensor data back to a remote operator. In some cases, the operator (or an associate) controls the vehicle; in others, the vehicle travels a pre-programmed path. Because RPVs are small and unobtrusive, they are ideal platforms for visual observation. Since the primary sensory information is visual, RPV applications stand to benefit strongly from the stereoscopic algorithms. The enhanced visual information would make the displays more realistic, and thus enable the operator to better interpret the RPV sensor's information. In addition, better visual fidelity would allow the RPV to perform some self-navigation, freeing up the operator for the more cognitively-demanding tasks such as analysis.

Possibly one of the most ambitious areas for the application of human-based stereopsis is in autonomous vehicles. These are mobile vehicles which operate without human intervention, picking a safe and viable path either on established roads or across country. These vehicles are programmed to start from a specific point, traverse an intermediate distance, and arrive at some destination. Only the beginning and endpoints are known. The vehicle relies on internal programming and problem-solving capabilities (artificial intelligence) to navigate to its goal. Since this stereopsis implementation is human-based, many of the problems inherent in other stereopsis methodologies (such as sensor noise, transparency, specular reflectances, lighting intensity changes, and ambiguous matches) would be minimized, or perhaps

even eliminated altogether. During a vehicle's traverse, it could perform any number of tasks that might require good depth information, including surveys, inspection, and intelligence gathering. Specific applications could include planetary exploration, land use surveys, factory and warehouse distribution systems, and patrol.

Other RPV/autonomous vehicle applications which could benefit from improved visual imagery include surveillance and image processing. In concert with pattern recognition, the procedures in this report could be useful in archaeology, resource management, land-use mapping, and military intelligence.

TELEOPERATION

Although it has much in common with RPVs, teleoperation has a wide range of applications which warrant separate discussion. Teleoperation is, in effect, the replacement of one visual reality with another. It is a method for placing a human into an environment which would otherwise be too hostile or expensive. Although crude implementations currently exist, most notably in the nuclear power industry, the remote manipulators used in many powerplants and nuclear fuel handling centers are rudimentary. The addition of high-quality imagery would vastly improve human performance, since it would more realistically place the human in the environment within which the task is being performed.

NASA is currently developing three-dimensional displays to present a robot's-eye-view to a human operator. This robot, designed to replace the human for routine exterior tasks in the space station, allows the operator to "see" what the robot sees. With the addition of remote manipulators, the human would be able to perform tasks in space from the shirtsleeve environment of the space station. In addition to orbital assembly and

nuclear powerplant operations, applications for work robots include oil exploration and drilling, firefighting, mining, and various military applications.

ROBOT VISION

In less dynamic (but no less demanding) environments, there are applications of the stereopsis algorithms to robotic assembly, quality control, inspection, and security systems.

MEDICAL APPLICATIONS

Many of the applications which have already been discussed could be applied to medical procedures as well. The algorithms could be used in medical imaging, or remote imaging for noninvasive surgery, such as arthroscopic procedures.

FUTURE DIRECTIONS AND ENHANCEMENTS

This phase I effort has spearheaded the development of an intelligent approach to the application of human cognitive and visual abilities to nonhuman computer-based systems. The human visual system has evolved to a remarkable state of competence. The application of this competence provides a solid foundation upon which to model an efficient and competent machine system.

TABLE OF FIGURES

REFERENCES

Arnold, R.D. (1982) Automated stereo perception. PhD thesis, Stanford University.

Baker, H.H. (1980) Edge-based stereo correlation. Proc. ARPA Image Understanding Workshop, Univ. Md., 168-76.

Baker, H.H. (1982) Depth from edge- and intensity-based stereo. PhD thesis, Univ Ill., Urbana, Ill.

Baker, H.H. & Binford, T.O. (1981) Depth from edge- and intensity-based stereo. 7th Intern. Joint. Conf. on Artif Intell. Vancouver, British Columbia, 631-36.

Barlow, H., Blakemore, C. & Pettigrew, J.D. (1967) The neural mechanism of binocular depth discrimination. J. Physiol. 193: 327-342.

Bough, E.W. (1970) Stereoscopic vision in the macaque monkey: A behavioural demonstration. Nature, 225: 42-44.

Cormack, R. & Menenedez, A. (1983) The role of convergence in stereoscopic depth constancy. Technical Report N14-0001-83C-0001.

Daugman, J.G. (1980) Two-dimensional spectral analysis of cortical receptive field profiles. Vision Research, 20: 847-856.

Davis, L. (1975) A survey of edge detection methods. CGVIP, 4: 248-70.

Dey, P. (1975) Perception of depth surfaces in random-dot

stereograms: A neural model. Int. J. Man-Machine Studies. 7: 511-528.

DeValois, A.L., Yund, E.W. & Helper, N. (1982) The orientation and direction selectivity of cells in macaque visual cortex. Vision Res., 22: 279-307.

Foley, J.M. (1980) Binocular distance perception. Psych. Rev. 87: 5, 411-434.

Foley, J.M. (1987) Interfaces for advanced computing. Scientific American, 257: 11, 127-135.

Foster, K.H., Gaska, J.P., Nagler, M. & Pollen, D.A. (1985) Spatial and temporal frequency selectivity in V1 and V2 of the macque monkey. J. Physiology, 365: 331-363.

Gabor, D. (1946) Theory of communication. JIEE, 93: 429-459.

Gennery, D.B. (1980) Modeling the environment of an exploring vehicle by means of stereo vision. Stanford Artif. Intell. Lab. Memo 339.

Grimson, W.E.L. (1981) From images to surfaces: A computational study of the human early visual system. Cambridge, Mass: MIT Press. 274 pp.

Julesz, B. (1960) Binocular depth perception of computer-generated patterns. Bell System Tech. J. 39: 1125-1162.

Julesz, B. (1971) Foundations of Cyclopean Perception. Chicago: Univ. Chicago Press. 406 pp.

Kass, M. (1983) A computational framework for the visual correspondence problem. Proc. ARPA Image Understanding

Workshop. Washington, DC.

Lu C. & Fender DH. (1972) The interaction of color and luminance in stereopsis.

Maffiei, L. & Fiorentini, A. (1973) The visual cortex as a spatial frequency analyzer. Vision Res. 13, 1255-1267.

Mallot, H & Bultoff, H. H. (1987) Stereo matching without zero crossings. Investigative Opthalmology and Visual Science (Supplement) 28, p364.

Marcelja, S. (1980) Mathematical description of the responses of simple cortical cells. Journal of the Optical Society of America, 70: 1297-1300.

Marr, D., Palm, G. & Poggio, T. (1977) Analysis of a cooperative stereo algorithm. Technical memo AI-M-446.

Marr, D., Palm, G. & Poggio, T. (1978) Analysis of a cooperative stereo algorithm. Biological Cybernetics, 28: 223-239.

Marr, D. & Poggio, T. (1976) Cooperative computation of stereo disparity. Science, 194: 283-287.

Marr, D. & Poggio, T. (1977) A theory of human stereovision. Technical memo AI-M-451.

Marr, D. & Poggio, T. (1979) A computational theory of human stereo vision. Proc. Royal Soc. London Ser. B. 204: 301-328.

Mayhew, J.E.W. & Frisby, J.P. (1980) The computation of binocular edges. Perception, 9: 69-86.

Mayhew, J.E.W. & Frisby, J.P. (1981) Psychophysical and computational studies towards a theory of human stereopsis. Artificial Intelligence, 16: 349-385.

Movshon, J. A., Thompson, I.D. & Tolhurst, D.J. (1978) Receptive field organization of complex cells in the cat's striate cortex. J. Physiol. 283, 79-99.

Nelson, J.I. (1975) Globality and stereoscopic fusion in binocular vision. J. Theoretical. Biol., 49: 1-88.

Nelson, J.I. (1977) The plasticity of correspondence: After-effects, illusion, and horopter shifts in depth perception. J. Theoretical Biol., 66: 203-266.

Ohzawa, I. & Freeman, R.D. (1986a) The binocular organization of simple cells in the cat's visual cortex. J. Neurophysiology, 56: 221-242.

Ohzawa, I. & Freeman, R.D. (1986b) The binocular organization of complex cells in the cat's visual cortex. J. Neurophysiology, 56: 243-259.

Palmer, L.A., Jones, J.P. & Mullikin, W.H. (1985) Functional organization of simple cell receptive fields. In: Models of Visual Cortex. Eds. D. Rose & V. Dobson. John Wiley & Sons, Chichester, 273-280.

Poggio, G.F. & Fisher, B. (1977) Binocular interaction and depth sensitivity of striate and prestriate cortical neurons of the behaving rhesus monkey. J. Neurophysiology, 40: 1392-1405.

Poggio, G.F. & Poggio, T. (1984) The analysis of stereopsis. Ann. Rev. Neuroscience, 7: 379-412.

Pollen, D.A. & Ronner, S.F. (1981) Phase relationships between adjacent simple cells in the visual cortex. Science, 212: 1409-1411.

Pollen, D.A., Gaska, J.P. & Jacobson, L.D. (1988) Responses of simple and complex cells to compound sine-wave gratings. Vision Res., 28(1): 25-39.

Scwartz, E.L. (1977) Spatial mapping in primate visual cortex: analytic structure and relevance to perception. Biolog. Cybernetics 25, 181-194.

Sperling, G. (1970) Binocular vision: A physical and a neural theory. Amer. J. Psychology, 83: 461-534.

ADDENDUM

During the Phase 1 effort, we have developed a new biologically motivated approach to stereoscopic depth estimation. In this approach, a pair of 4-D Gabor representations is computed in each of many "disparity channels". For each channel, the associated pair of Gabor representations are cross-correlated and pooled to produce a 2-D normalized cross-correlation function. When this function is combined with the correlation results of all other disparity channels, there results a 3-D normalized depth correlation function from which a depth map is readily obtained by selecting, for each direction in visual space, the disparity at which this correlation function is maximized. A complete description of the full complexity of this new approach is discussed in detail in the previously submitted Phase 1 final report.

During the Phase 1 effort, we also developed a restricted implementation of the new approach which enabled us to offer a practical demonstration of its feasibility. The purpose of such an implementation was to establish a proof of concept for the developed disparity algorithms. For the imagery on which to conduct this demonstration, we chose to employ random-dot stereograms to test the current implementation of the algorithm. Such stereograms are frequently (nearly universally) used in initial tests of new stereo-depth estimation algorithms as they are considered as the most appropriate imagery in terms of providing a critical disparity estimation test. In particular, as described in the final report, computational limitations of the systems current implementation made it unsuitable for processing real-world (natural-camera) images in a test of feasibility.

First, as will be further elaborated in this addendum, the new approach is, in general, very computationally demanding and

thus imposes severe constraints on the capability to fully process complex analog-in-depth imagery.

Second, due to the short 6-month duration of a Phase 1 effort. it was only possible to develop a restricted system capable of demonstrating the essential elements of the new approach in terms of feasibility. In particular, as was already clearly elaborated in the final report, the current implementation has two algorithmic limitations which severely restrict its application to analog-in-depth imagery: (a) the inability to pool over multiple spatial-frequency scales at each point in the visual field, and (b) the inability to process multiple (greater than two) planes of disparities. We review these limitations below.

The current restriction to pooling over a single spatial-frequency scale at each eccentricity limits the effectiveness of the implemented system at handling imagery that do not everywhere have broadband spectral characteristics. However this does not limit the ability to demonstrate feasibility as imagery with such broadband characteristics are readily available. Given this limitation, random-dot stereograms are ideal for demonstrating the approach in its current implementation. This is due to the fact that in addition to providing a salient and stringent testing medium, random dot stereograms, by virtue of their pointillistic patterning are spectrally broadband. Such stereo image pairs were therefore used to demonstrate feasibility of the new approach in the Phase 1 final report.

The restriction of the current implementation to two disparity channels causes the current system to assign each point in the visual field to one of two possible depth planes. At this level of implementation the existing system is not appropriate for processing stereo imagery that depicts a scene containing

complex 3-dimensional objects whose surfaces lie at many depths relative to the observer (i.e. analog-in-depth). For this reason we used, in the final report, a "pure" stereogram structured such that each element had residence in one of two unique depth planes in order to demonstrate the feasibility of the new approach to stereo depth estimation.

Importantly, the above restrictions do not apply to our new approach, but apply only to the current restricted implementation devised for the purpose of proof of concept given the nature of the limitations imposed by the inherent computational complexity of the task. The current implementation was solely developed to evaluate the potential of the approach. The level of effort required to fully implement the proposed new approach in a prototype hardware/software system that is designed to robustly handle real-world analog-in-depth imagery without restriction would constitute a major effort over at least a two-year period.

Therefore, due to computing facility and algorithmic limitations of the current implementation, we felt that it would not be very informative, in terms of proof of feasibility, to process real-world (camera) images for inclusion in the Phase 1 final report. Nonetheless, we originally proposed in the Phase 1 proposal to process real-world imagery during the feasibility study. This was proposed before the details (as well as some of the major physiologically-based processing strategies) of the new approach had been invented, and hence it was not possible to anticipate the high computational burden of the developed approach relative to the capabilities of available computing capacity.

This addendum describes the processing of a real-world (camera) stereo-image pair to fulfill our obligation to process such a image as delineated in the Phase 1 proposal. The decision/notification of the necessity to demonstrate the new

approach on real-world imagery occurred after submittal of the Phase 1 final report. Thus, the documentation of the results of such testing are reported in this addendum. We as researchers originally had serious reservations as to the appropriateness of such a test in the respect that the model implementation was designed for proof of feasibility and was not implemented in the potential full level of complexity which we felt would be necessary to handle real-world analog-in-depth imagery. Despite such reservations, a mutual agreement was reached in which one real-world (camera) stereo image pair would be processed using the current restricted implementation of the new depth estimation approach. Such a test has subsequently been performed and its outcome is discussed below.

Real-world (Camera) Image Example

The criteria we employed in our choice of a real-world analog-in-depth view from which to obtain a stereo image pair for processing required a salient complex scene structure with multiple depth planes and multiple rates of transition across the depth planes. We chose a computer board, in particular an EGA video graphics controller board, as the real-world scene. This produces a complex analog-in-depth image with many planar as well as slanted surfaces. We used a consumer-grade RCA Neuvicon camera to directly produce digitized stereo images of the EGA video graphics controller board. The two differing views of the object were produced by laterally displacing the camera whose optical axis was maintained normal to the surface of the circuit board. From each of these digitized images, we extracted a 265 by 265 pixel subimage such that these subimages depicted the same general region of the circuit board. The resulting stereo image pair is depicted as the stereogram of Figure A1 (a) and (b). The stereo images produce a rich depth percept with the proper ordering of surfaces in depth when convergently fused by the reader (i.e., A1(a) should be viewed by the right eye and A1(b)

should be viewed by the left eye.)

In the 265 by 265 pixel coordinate frame of each image in this stereo pair, the planar surface region on the large PROM chip with the writing "SEEQ" is in 0-pixel disparity relation, and the other visible surfaces all appear further away (when convergently viewed) and have positive disparity relations ranging up to approximately 15 pixels. In other words the planar surface of this circuit chip is the "highest" surface in the image. The integrated circuit portion of this PROM chip, which is visible through the circular window in the center of the chip, has a 2-pixel disparity relation in the stereo image pair and thus appears to be noticeably below the planar surface of the chip. For convenience a schematic map of disparities associated with different surfaces is depicted in Figure A2. These disparities were measured by inspection of the interference images produced by shifting and subtracting the left and right eye images in the stereo pair.

We chose the two disparity channels in the implemented depth estimation system as 0 and 2 pixels disparity. We wished therefore to test the ability of the existing implementation of the system to correctly classify surfaces appearing at 0 and 2 pixels disparity. In other words, if the model performed ideally it should be capable of segregating those two regions of the image, namely, the top of the large chip (residing at 0 disparity) and the IC circuit and surrounding bonding pad region (residing at 2 pixels disparity). Surfaces which appear in the images at other disparities not processed by the system were expected to be incorrectly and randomly assigned to either the 0 or 2 pixel disparity plane.

The interesting solution obtained using the existing implementation is shown in Figure A3 (a) and (b) where the system solution is depicted by the colorization process of the original

stereo image pair. Red regions were estimated to have a 0-pixel disparity relation and green regions were estimated to have a 2-pixel disparity relation. This solution can therefore be directly compared by the reader to the "correct" disparity map (the percept produced through convergent fusion) as already shown in Figure A2. Recall that this map was manually determined as well before conducting the experiment. Note that, due to the foveated nature of the implemented system, colorized disparity estimates are only available within a circular region whose diameter equals the width of the images. The outer corners of the stereo images, outside the central circular region, have been arbitrarily colorized green.

Note that the leftward region of the large PROM chip that contains the writing is, for the most part, correctly classified as having 0 disparity (red) but that the disparity classification becomes more random as one progresses to the opposite (right) end of the PROM chip. Similarly, the IC circuit and bonding pad region visible through the circular window in the large PROM is largely classified correctly at 2 pixels disparity (green). Note, however, that the boundary between these two regions (where the analog transition across the two depth plane occurs) is rather poorly demarcated.

There are several possible explanations for the relatively random classification results on the right side of the PROM chip. One possibility is that the PROM chip may actually lie slightly further away from the camera on its right side than the left since the camera's optical axis may have been slightly tilted away from perfect normalcy. Note that we obtained manual measurements of the disparity only on the left half of the PROM chip where the appearance of writing facilitated this task. Hence the disparity may be intermediate between 0 and 2 pixels of disparity, leading to the production of this rather random classification. This would not be a possible problem in a future

implementation of the approach which employs a dense array of disparity channels. Alternatively, the sparsity of conspicuous luminance variations on the right side of the PROM (as is produced by the printed text on the left side of the chip) may be causing problems for the implemented system. This would be especially problematic since the existing implementation pools information from filters at a single spatial-frequency scale at each point in the visual field. A future complete implementation of the new approach should show substantially improved performance by virtue of its ability to exploit subtle variations in the left and right eye Gabor image representations at both high and low spatial frequencies.

The inability of the existing system to pool over multiple spatial frequencies may also be responsible for the relatively poor disparity classification results at the circular boundary between the ceramic PROM chip surface and the bonding pad region visible through the window in the PROM.

Other regions of the image residing at various disparity planes are randomly classified, as expected, since their disparities are not represented in the present system.

In summary, the performance of the current implementation of the model on an analog-in-depth stereo image was encouraging. For disparity planes outside of either the 0-pixel or 2-pixel disparity planes the depth assignments of the model were essentially random, as they should be. The "ragged" results at the transition across the two disparity planes as well as the misclassification of significant portions of the right side of the PROM chip is demonstrative of the necessity for multiple spatial scales as well as a more dense sampling of disparity planes which the fully implemented model would possess. Finally, the misclassification of portions on the right side of the PROM chip is indicative of the sensitivity to minor depth changes

and/or spatial scale changes that a single disparity channel possesses.

Complexity Analysis

This section offers a brief discussion of the computational requirements of a hypothetical system which generalizes the existing implementation by pooling across a range of spatial frequencies at each eccentricity, and by permitting one to incorporate a user-specified number of disparity channels.

Consider the stereo image pair of the circuit board which was used in the experiments which we have just discussed above. This image pair has a total disparity range of about 15 pixels. Hence about 15 depth channels would be required to fully test the new depth estimation approach on this image. As discussed in the previously submitted Phase 1 final report, the newly developed approach to depth estimation requires that 2 complete Gabor functions must be computed within each disparity channel. Therefore, for the stereo images of the circuit board, a total of about 30 Gabor representations must be computed to process all 15 depth planes. In the current restricted implementation of the approach, the computation and storage complexity associated with the computation of each Gabor function in the previous circuit board example has been empirically determined to be as follows:

To compute 1 Gabor representation over:

- eccentricities from 0.075 to 10 degrees.
- spatial frequencies from 0.05 to 6.0 cycles/deg.
- 6 frequency scales (levels).
- 6 frequency orientations (phases).

requires:

- 1.5 MB storage for conformally mapped lowpass image pyramid.
- 8.0 MB storage for Gabor representation.

AND

- 1875 MFLOP (million floating point operations) for space

variant filtering of the image to obtain lowpass pyramid.
600 MFLOP for space-invariant Gabor filtering of lowpass pyramid input.
Totalling 2475 MFLOP per computed Gabor representation.

Once the Gabor representations have been computed from the original stereoscopic images, the additional computational demands associated with Gabor correlation, pooling and depth discrimination are relatively minor. Hence, the computational demands of the new disparity estimation approach are well approximated by those associated with the front-end linear filtering operations alone!

As an approximate rule of thumb, each doubling of the high frequency range over which the Gabor is computed leads to a doubling of storage and computation requirements. Similarly, each doubling of the range of eccentricities over which the Gabor is computed leads to increases in storage and computation by a factor of about (n+1)/n where n is the range in octaves and

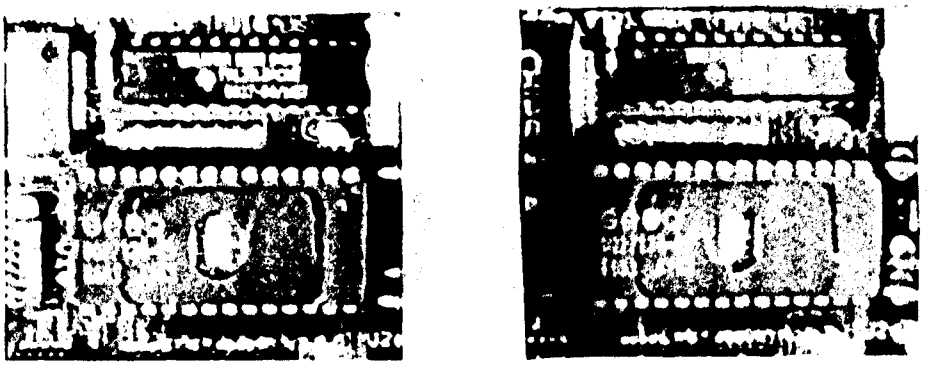
$$n = \log_2 (\text{eccentricity}_{hi} / \text{eccentricity}_{low}),$$

from the lowest to the highest represented eccentricity before doubling the range. It is therefore reasonably expensive to double the peak resolution of the system, but, because of the space-variant (foveated log conformal mapping) nature of the encoded representation as discussed in the Phase 1 final report, increasing the visual field extent is relatively inexpensive computationally.

Given the computational complexity outlined above, it is easy to determine the cost of computing the 30 Gabor level representations necessary to fully process all disparities in the circuit board stereogram. Namely, 30 Gabors x 2475 MFLOP/Gabor = 75,000 MFLOP (approximately). This would require about 52 days of CPU time on a 10 MHz PC/AT with an 80287 which we currently have available to us. A 20 MFLOP/sec array processor would bring

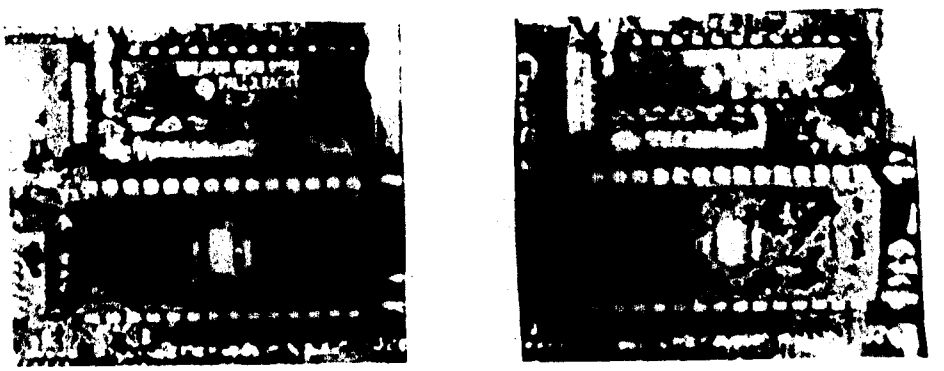
this time down to a couple of hours. In a production system, some parametric flexibility of the algorithms could be sacrificed in favor of speed to get about a factor of 2 to 5 reduction in the required FLOPS.

In a system designed to handle any analog real-world images, one would employ disparity channels of fine resolution near the depth of fixation (zero disparity) and progressively more coarse disparity sampling at increasing distances in front of and behind the plane of fixation. This would lead to a reduction in overall computational cost without sacrificing either resolution near the plane of fixation, or range of total disparity encoding. These additional developments would further reduce the computational complexity of the approach and would be an integral part of any subsequent phase of development.



(a) (b)

Figure A1. Real-world stereogram produced from camera images of a section of an EGA video graphics controller board. The salient depth percept can be produced by the reader by convergent fusion of the image pair.
(a) Right-eye image. (b) Left-eye image.



(a) (b)

Figure A3. Colorized version of the real-world stereogram depicted in figure A1. Colorization represents the models solution to the analysis of two disparity planes. The models estimate of the 0-pixel disparity plane is represented in red and the 2-pixel disparity plane is represented in green.
(a) Right-eye image. (b) Left-eye image.

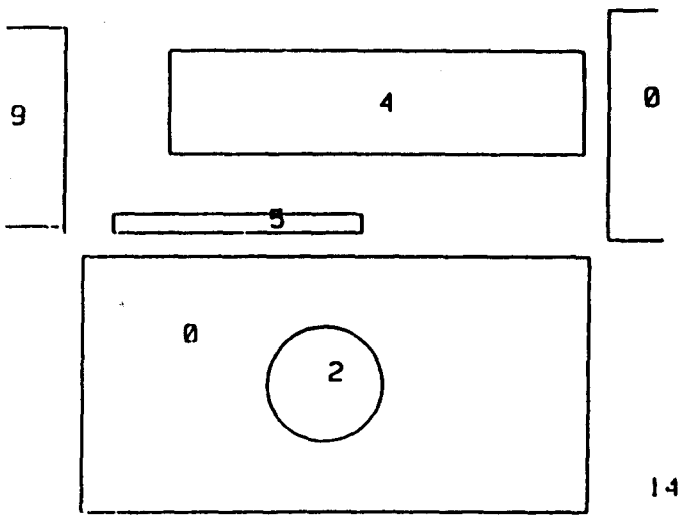


Figure A2. A schematic map of the disparities associated with different surfaces depicted in figures A1 and A3. The numbers represent the disparity values in pixel units determined from the interference images of the stereo pair in figure A1.

* U.S. GOVERNMENT PRINTING OFFICE: 1988 • 548-002/80208